Supplementary Figures, Notes, and Tables to the manuscript i-cisTarget: an integrative genomics method for the prediction of regulatory features and *cis*-regulatory modules

Carl Herrmann[1,2,*], Bram Van de Sande[3,*], Delphine Potier[1,3], and Stein Aerts[3]

1 TAGC – Inserm U928 and Aix-Marseille Université, Campus de Luminy, Marseille, France

2 Département de Biologie, Campus de Luminy, Aix-Marseille Université, Marseille, France

3 Laboratory of Computational Biology, Center for Human Genetics, Leuven, Belgium

***** These authors contributed equally to this work

Note S1: Possible explanations for failures in motif-enrichment analyses

For two TFs, namely eyeless and prospero, the failure to detect the expected motif might be explained by the specificity of the motif to certain conditions. Therefore, we investigated the gene sets for which i-cisTarget apparently failed to identify the correct motif, i.e. the ey-GOF and pros-LOF sets. Concerning eyeless, we looked at another co-expressed gene set, namely genes expressed in the mushroom body (MB) according to FlyBase. The MB is a structure in the *Drosophila* brain involved in learning and memory where *ey* is also involved (1). In that set we easily find the *ey* PWM. Therefore, the ey binding site in the eye could be different from the ey binding site in the mushroom body, and the available PWM for ey (based on in vitro binding specificities) may rather reflect the specificity in the mushroom body. Concerning prospero, we found that the pros motifs in our library (one from the SelexConsensus set (CWNNNCY) and one derived from a consensus site (CWYBDCY)) are very different from the pros motif reported by Choksi et al based on a DamID experiment (TWAGNY) (2). Moreover, these pros motifs are also not detected directly on the DamID peaks (data not shown; we describe further below how i-cisTarget is used on peak regions). Therefore, the reason for not finding the pros motif is rather a biological problem of the dataset, than a failure of i-cisTarget.

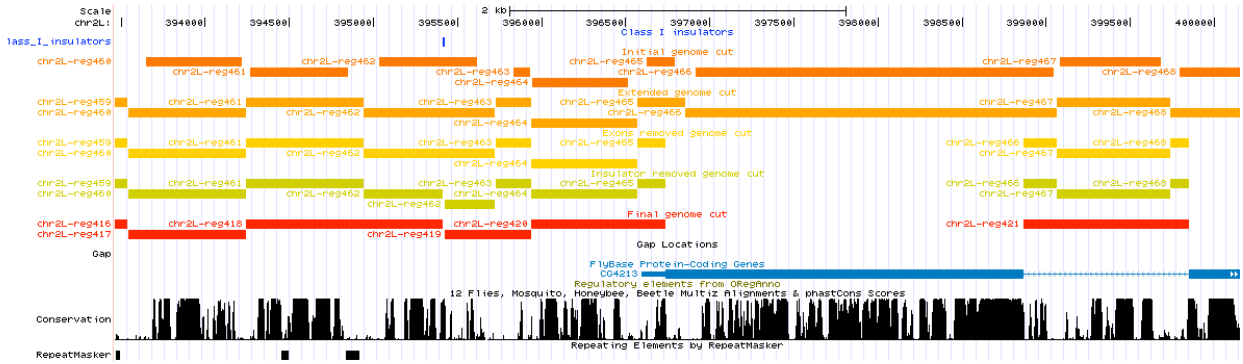


Figure S1. Partitioning the non-coding genome

UCSC genome browser screenshot illustrating the different steps in the genome cutting procedure: Definition of initial seeds around peaks of conservation (orange track); Extension to obtain a full coverage of the genome (light orange track); CDS removal (yellow track); Insulator split (green track); Extension to a minimal length of 500 bp (red track).

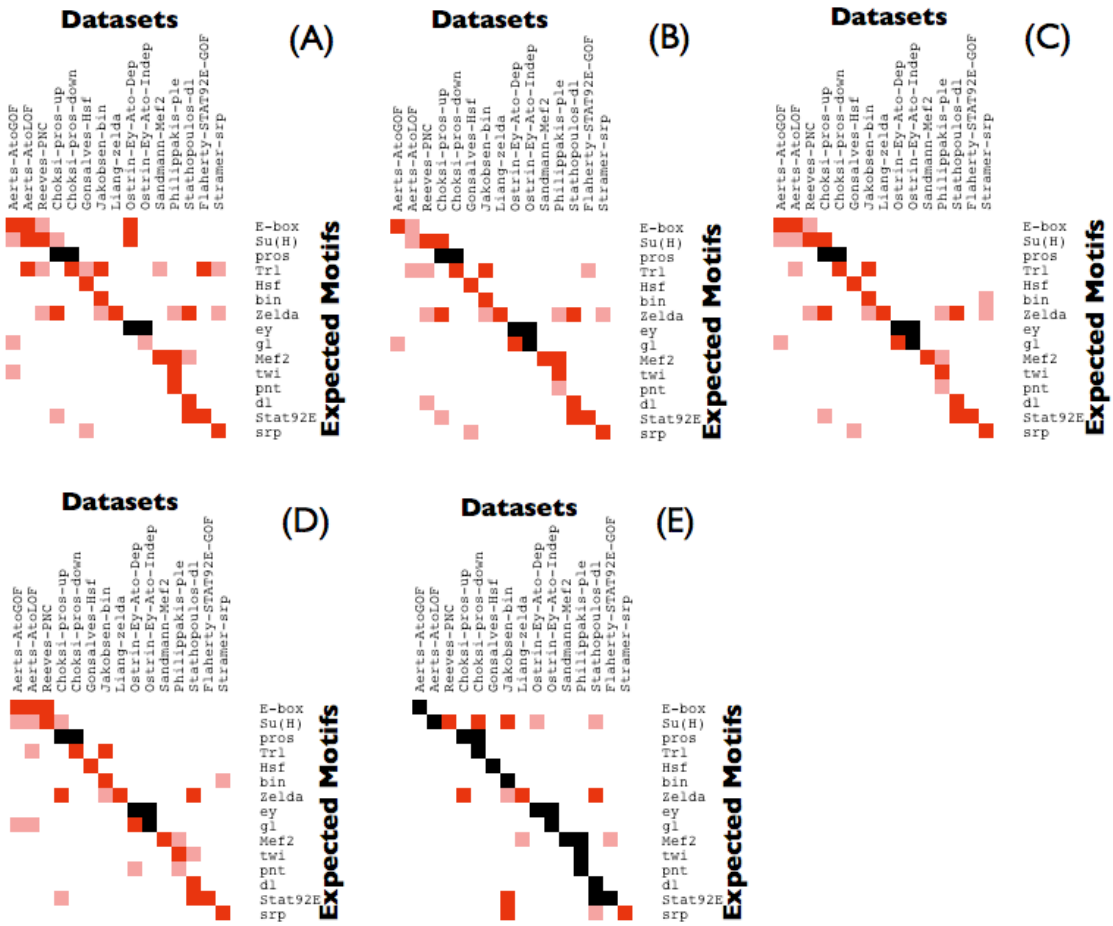


Figure S2. Motif discovery performance for different putative regulatory region demarcations used for mapping gene signatures to the 136K scored regions

Heatmaps displaying the motifs discovered in various gene sets datasets; red indicates that the motifs ranks among the top 3 motifs, pink that the motif has an enrichment score above the NES threshold (NES >= 2.5), and black indicates that the expected motif is not found (The same color code used as in figure 3B). Motifs discovered when using the default 5kb upstream, 5'UTR and first intron demarcation (A), 5kb upstream full transcript demarcation (B), 5kb upstream, full transcript and 5kb downstream (C), 10kb upstream, full transcripts and 10kb downstream (D). All these demarcations exclude coding sequences. The last heatmap shows the discovered motifs when using a 100kb upstream of the TSS + 100kb downstream of the TSS demarcation not limited by nearest genes and without removal of coding sequences (E).

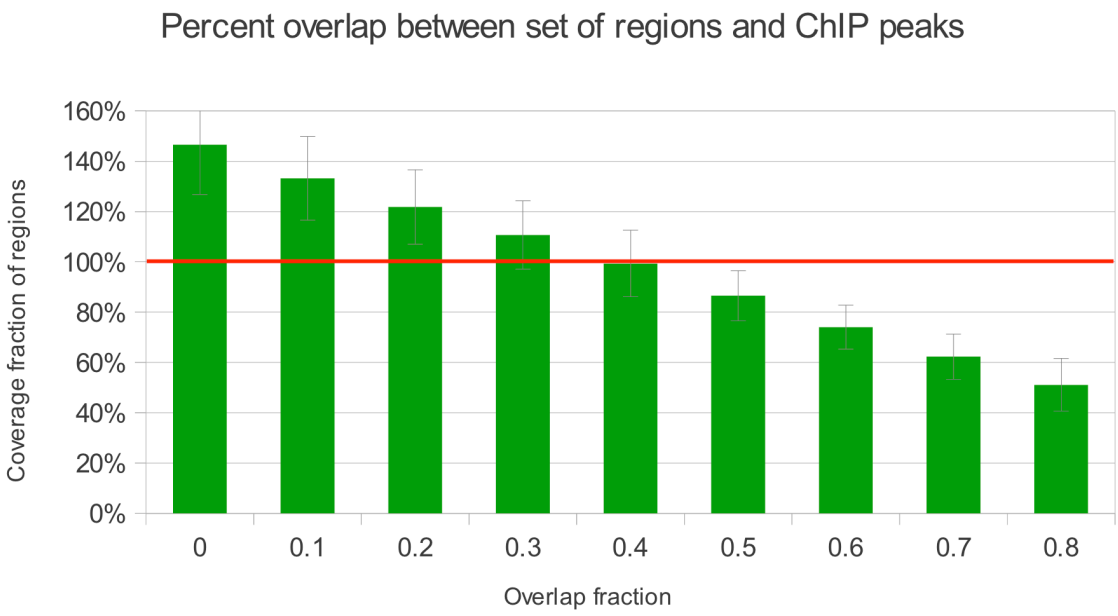


Figure S3. Overlap between ChIP peaks and predefined regions.

Histogram showing the coverage fraction of the sets of regions for the 40 BDTNP ChIP datasets. We vary the overlap fraction (-f option in intersectBed), determine the size of the regions obtained, and compare it to the size of the original ChIP peaks. Small overlap parameter result in a set of regions that exceeds the size of the original dataset, while a stringent overlap (close to 1) parameter results in a set of regions that does not cover the ChIP peaks. The optimal parameter appears to be -f 0.4, resulting in an average coverage of 99.35%.

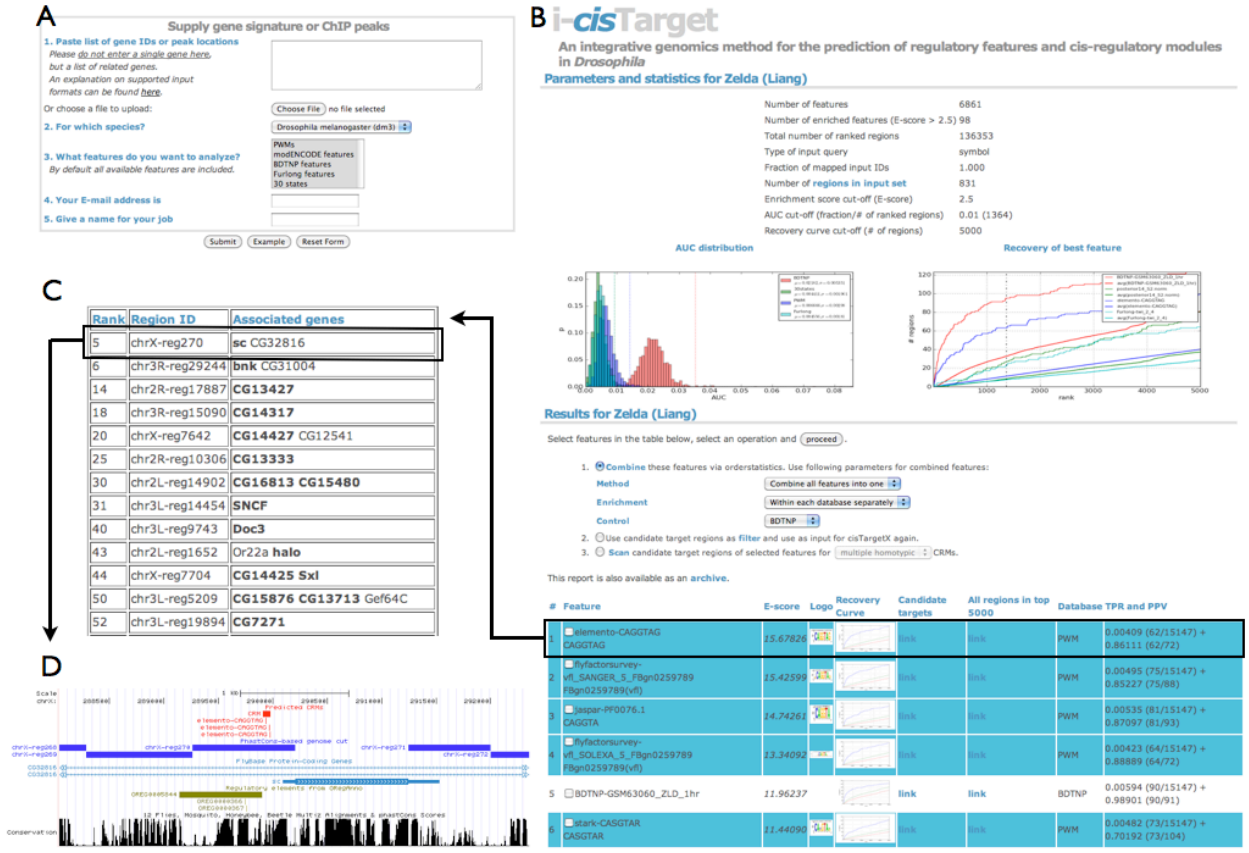
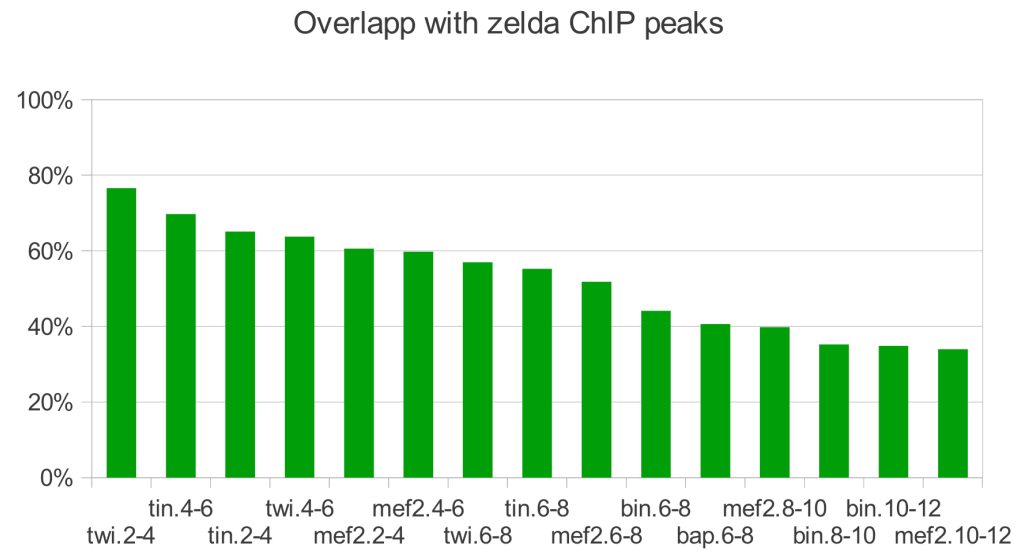


Figure S4. Screenshots and workflow of i-cisTarget.

(A) The input form on the i-cisTarget web interface. We used the Liang (2008) zelda LOF microarray signature as input (3). (B) The report providing an overview of enriched features. Note that i-cisTarget correctly finds the zelda motif ranked first, several other Zelda motif variations (all in blue), and the Zelda ChIP dataset as the best iVE. Check boxes before each feature allow to select this feature for combinatorial analysis. (C) i-cisTarget also supplies an overview of the regions that are considered candidate enhancers for this feature and also of the genes for which the associated regulatory regions intersects with these candidate enhancers. The genes that are part of the input gene signature are indicated in bold. (D) The actual predicted CRMs and their constituent motifs, located in these candidate enhancers, can be investigated in the UCSC genome browser.

(A)



(B)

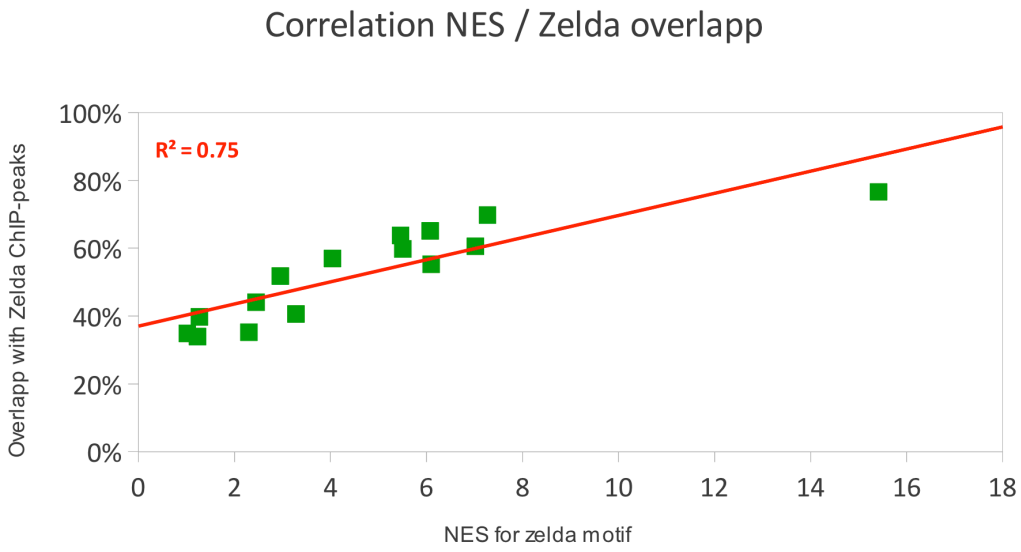


Figure S5. Comparison of the Zelda motif discovery with Zelda ChIP peaks

(A) Histogram showing the overlap between binding location for the five mesodermal transcription factors at different stages (4), and the zelda binding locations (5); (B) Correlation between the NES for the zelda motif in the different mesodermal datasets, and the actual overlapp with true binding locations, as in (A).

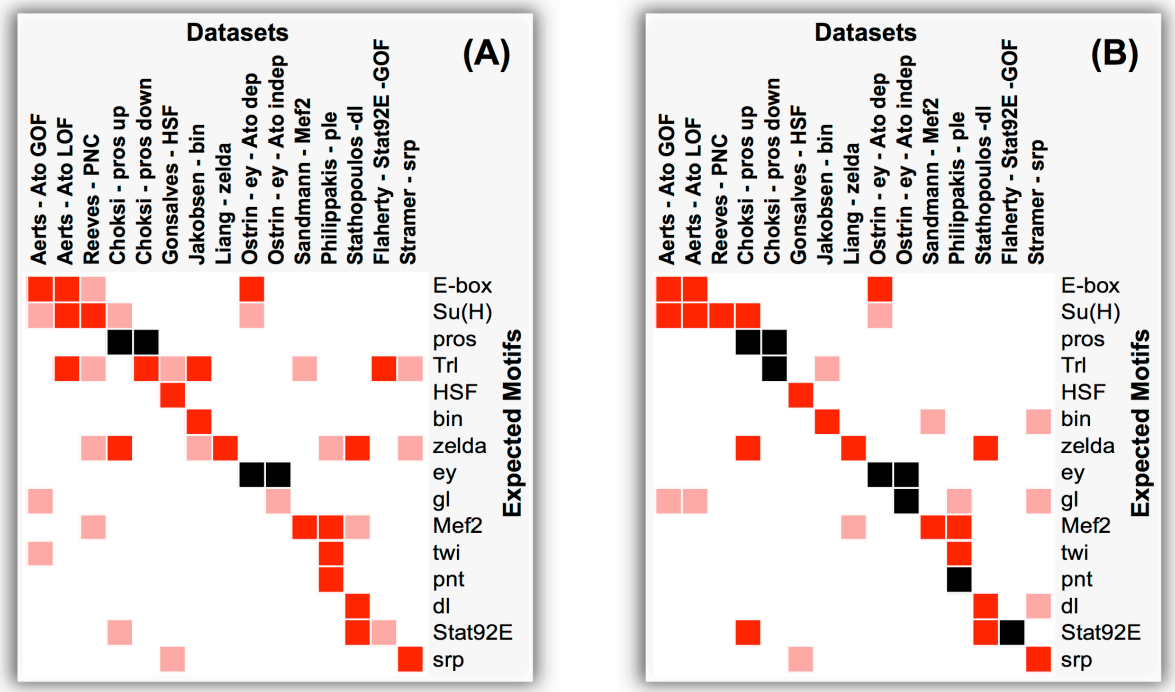


Figure S6. Comparison of i-cisTarget results in a gene-based or region-based approach.

(A) Region-based; this is the same figure as figure 3B in the main text; duplicated here for comparison. (B) The gene-based approach, as used in (6) has slightly more black squares, meaning that the motif corresponding to the expected TF was not identified as enriched, such as for Trl, gl, pnt, and Stat92E.

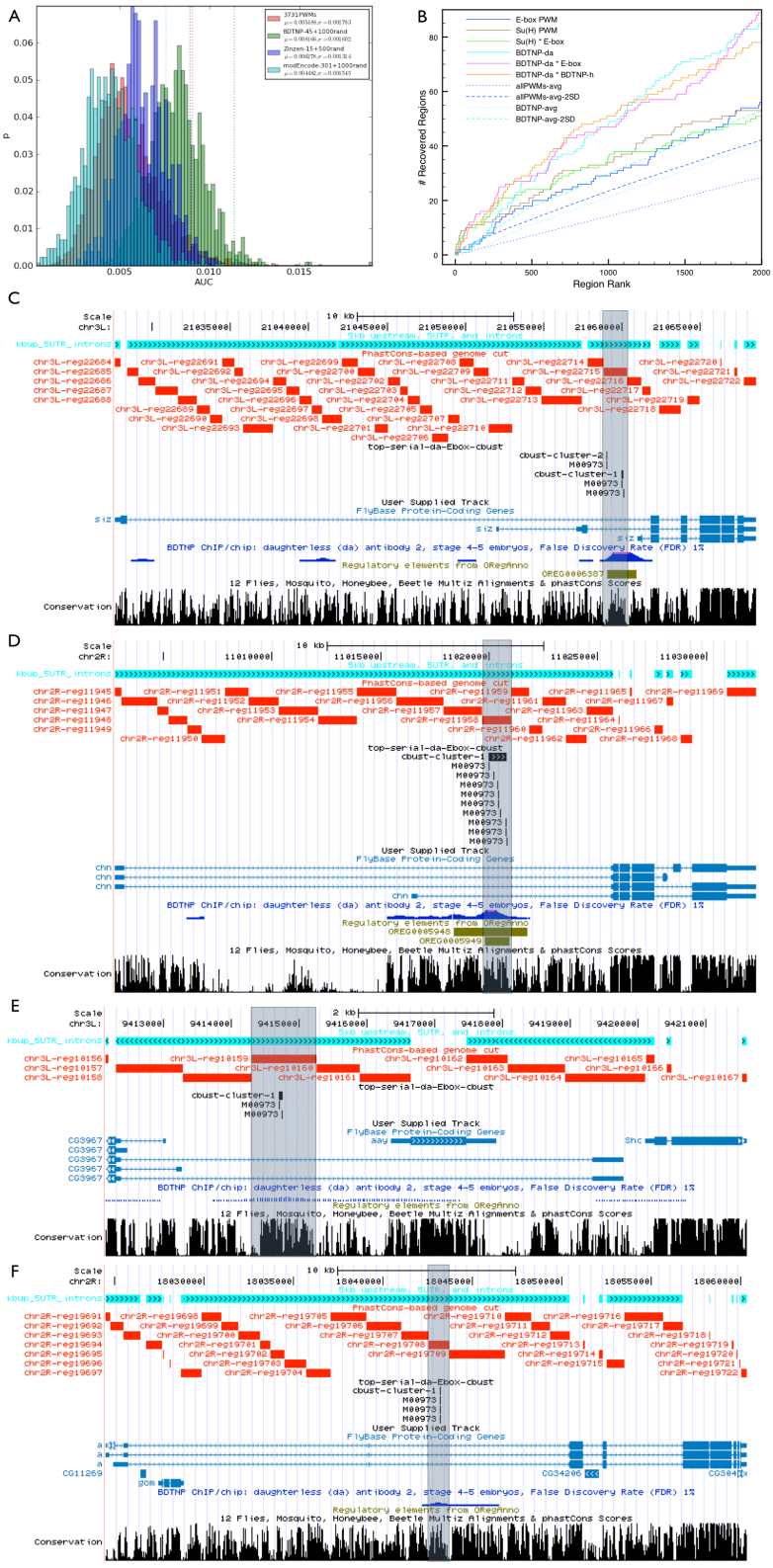

A
B
C
D
E
F
AUC
p
Recovered Regions
Region Rank
E-box PWM
Su(H) PWM
Su(H) * E-box
BDTNP-da
BDTNP-da * E-box
BDTNP-da * BDTNP-h
allPWMs-avg
allPWMs-avg-2SD
BDTNP-avg
BDTNP-avg-2SD
Scale
chr3L:
chr2R:
10 kb
kbup_SUTR_introns
5kb upstream, SUTR, and introns
PhastCons-based genome cut
top-serial-da-Ebox-cbust
cbust-cluster-1
cbust-cluster-2
User Supplied Track
FlyBase Protein-Coding Genes
BDTNP ChIP/chip: daughterless (da) antibody 2, stage 4-5 embryos, False Discovery Rate (FDR) 1%
Regulatory elements from OReganno
12 Flies, Mosquito, Honeybee, Beetle Multiz Alignments & phastCons Scores
Conservation

Figure S7. Motif and iVE discovery in PNC genes (figure on previous page)

Example case study on a set of co-expressed genes in proneural clusters in the wing imaginal disc (7). (A) Overlapping yet different AUC-1% distributions of different feature databases, including PWMs (pink), BDTNP features (green), Mesodermal features (blue), and modENCODE features (light blue). Enriched features for each database are calculated by a normalized enrichment score within each database separately. (B) Recovery curves for several top enriched features, including single motif features (the E-box and Su(H)), single in vivo features (daughterless ChIP data from BDTNP), combined PWMs (Ebox + Su(H)), combined iVEs (daughterless + hairy ChIP data), and combined PWM and iVE (Ebox + daughterless ChIP data). (C-F) Screenshots from the UCSC Genome Browser showing predicted CRMs using the combination of the Ebox PWM feature and the daughterless ChIP feature. C and D are predicted CRMs that overlap known CRMs (indicated by the ORegAnno track in green). E and F are examples of new CRM predictions.

Table S1: Supplementary Excel file with all the data sets used for validation (ChIP data sets and gene expression datasets), and all the individual motifs and iVEs.

References

1. Kurusu,M., Nagao,T., Walldorf,U., Flister,S., Gehring,W.J. and Furukubo-Tokunaga,K. (2000) Genetic control of development of the mushroom bodies, the associative learning centers in the Drosophila brain, by the eyeless, twin of eyeless, and Dachshund genes. *Proc Natl Acad Sci USA*, **97**, 2140–2144.

2. Choksi,S.P., Southall,T.D., Bossing,T., Edoff,K., de Wit,E., Fischer,B.E., van Steensel,B., Micklem,G. and Brand,A.H. (2006) Prospero acts as a binary switch between self-renewal and differentiation in Drosophila neural stem cells. *Dev Cell*, **11**, 775–789.

3. Liang,H.-L., Nien,C.-Y., Liu,H.-Y., Metzstein,M.M., Kirov,N. and Rushlow,C. (2008) The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. *Nature*, **456**, 400–403.

4. Zinzen,R.P., Girardot,C., Gagneur,J., Braun,M. and Furlong,E.E.M. (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, **462**, 65–70.

5. Harrison,M.M., Li,X.-Y., Kaplan,T., Botchan,M.R. and Eisen,M.B. (2011) Zelda Binding in the Early Drosophila melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition. *PLoS Genet*, **7**, e1002266.

6. Aerts,S., Quan,X.-J., Claeys,A., Sanchez,M.N., Tate,P., Yan,J. and Hassan,B.A. (2010) Robust Target Gene Discovery through Transcriptome Perturbations and Genome-Wide Enhancer Predictions in Drosophila Uncovers a Regulatory Basis for Sensory Specification. *PLoS Biol*, **8**, e1000435.

7. Reeves,N. and Posakony,J.W. (2005) Genetic programs activated by proneural proteins in the developing Drosophila PNS. *Dev Cell*, **8**, 413–425.